nag®

Combining Open MP and MPI within GLOMAP Mode: An Example of Legacy Software Keeping Pace with Hardware Developments

Mark Richardson, NAG

May 2010

Introduction

- What is GLOMAP?
- Who Funded the work?
- Goals of the project
- Background to the project
- Case description
- Analysis of the application prior to modification
- Results from XT4 and surprise-surprise XT6
- Summary
- Conclusions

What is GLOMAP?

- A computer program for simulating aerosol processes in the earth's atmosphere
  - TOMCAT an advection code is the main program
    - Reads wind data and transports the chemistry around the atmosphere
    - Maintained and Supplied by Professor Martyn Chipperfield, University of Leeds
  - GLOMAP Mode, the aerosol process method
    - Replaces the built-in chemistry model of TOMCAT (the subroutine "CHIMIE")
    - Developed at University of Leeds by Dr. Graham Mann, NCAS
  - ASAD the chemical reaction solver.
    - Dr. Glenn Carver, University of Cambridge

Acknowledgements

- NERC, NCAS
- Research Councils UK, HECToR Resource
- University of Leeds School of Earth and Environment
- HECToR DCSE programme
- NAG provide personnel (me)
- Additional technical support
  - HECToR Service Helpdesk
  - Cray Centre of Excellence
  - Portland Group (on-line forum)

Goals of the project

- Enhance the existing MPI version of GLOMAP MODE with Open MP directives
- Enable the mixed-mode of parallel operation for better use of the multi-core systems that are being installed
- Allow higher resolution simulations

Background

- MPI version has already been subject of a DCSE to enhance it's performance on the XT4h
- A solely Open MP version had already been developed prior to the MPI version
  - Several years ago and completely separate from this project.
  - Significant to this research
- Expect the inauguration of XT6 with 24 cores per node at Edinburgh within a few months.

Case description 1

- The T42 model is quite a low resolution model
  - 128 x 64 x 31 i.e. 2.8^{o} per grid-box
- The incorporation of chemistry makes it quite memory hungry
  - minimum 4 nodes of XT4h using 8 cores
  - i.e. hector phase 1, ~3GB per core
- The MPI version decomposes by two dimensions (lon by lat) but retains full altitude on each “patch”
- Each hexahedral division is called a grid-box
- Typical use is 32 MPI tasks (32 x 8 x 31)

Case description 2

- Earth's atmosphere mapped into a 3-D Cartesian coordinate system
- MPI 2-D topology creates uniform patches

- T42 is 128x64x31
- 197 scalars
- 3 days used for investigation
- (144 steps)
- I/O stages omitted as they form ~30% of simulation

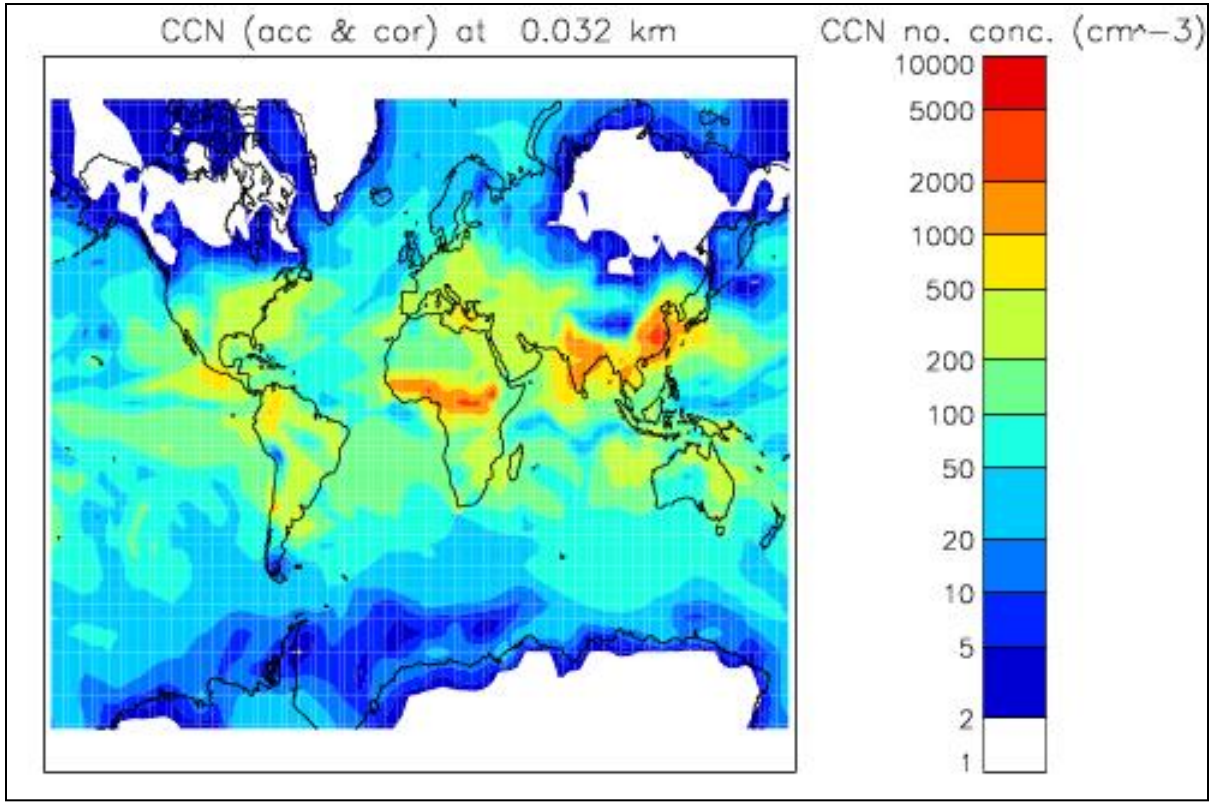


Table 2: selected Cray PAT results for fully populated nodes GLOMAP mode pure MPI version and three configurations

	M32 % of whole sim	M32 % of GMM only	M64 % of whole sim	M64 % of GMM only	M128 % of whole sim	M128 % of GMM only
ADVX2	4.2	5.7	2.8	5.5	1.3	5.7
ADVY2	11	14.9	7.1	13.9	2.2	9.7
ADVZ2	4.9	6.6	3.2	6.3	1.4	6.2
CONSOM	5.4	7.3	3.6	7.1	1.1	4.8
CHIMIE	40.9	55.3	27.4	53.7	12.4	54.6
MAIN	7.7	10.4	7	13.7	4.4	19.4
TOTAL FOR GMM	74	100	51	100	22.7	100
MPI	13.3	-	28.3	-	47.4	-
MPI_SYNC	12.7	-	20.7	-	29.9	-

TOMCAT Analysis

- 4 subroutines in TOMCAT accounts for ~30% of the GMM
- ADVX2 an outermost loop over NIV
  - Upper limit 31 at the test case resolution
- ADVY2 an outer most loop over NIV
  - Upper limit 31 at the test case resolution
  - Has some extra MPI work for polar regions
- ADVZ2 an outermost loop over MYLAT
  - Upper limit (16,8,4,2) for (16,32,64,128) tasks
- CONSOM a second level loop over NTRA
  - Upper limit 36
  - The outer loop is over MYLAT

CHIMIE Analysis

- The CHIMIE subroutine accounts for ~55 %
  - has some MPI work and additional loops external to major loop
- Contains a major loop over latitudes (MYLAT)
  - Upper limit (16,8,4,2) for (16,32,64,128) tasks
- THREADPRIVATE common block for interfacing with ASAD
- Large block of code explicitly declaring
  - Private data passed into the GLOMAP sub-system
  - Shared data
  - Common block

ASAD Analysis

- ASAD had already been converted for use with the earlier version of GLOMAP with Open MP
  - This is an ODE solver for chemical reactions
  - It is wholly within the CHIMIE subroutine
- Only common blocks have been treated to retain private data
- No “acceleration” from Open MP explicitly
  - The time spent within ASAD is spread between the threads
  - Time for computation in the MPI task is reduced
- Possible future project on balancing chemistry calculations

Systems

- XT4h
  - Initially dual-core Opteron
    - 2 cores per node and ~3GB per core
  - Now with a single quad-core “Barcelona”
    - 4 cores per node and ~2GB per core
- XT6
  - Two “Magny-Cours” processors
    - 24 cores per node ; ~1.5GB per core
  - Specific description will appear at:
    - http://www.hector.ac.uk

XT6 Nodal arrangement: crude schematic on purpose

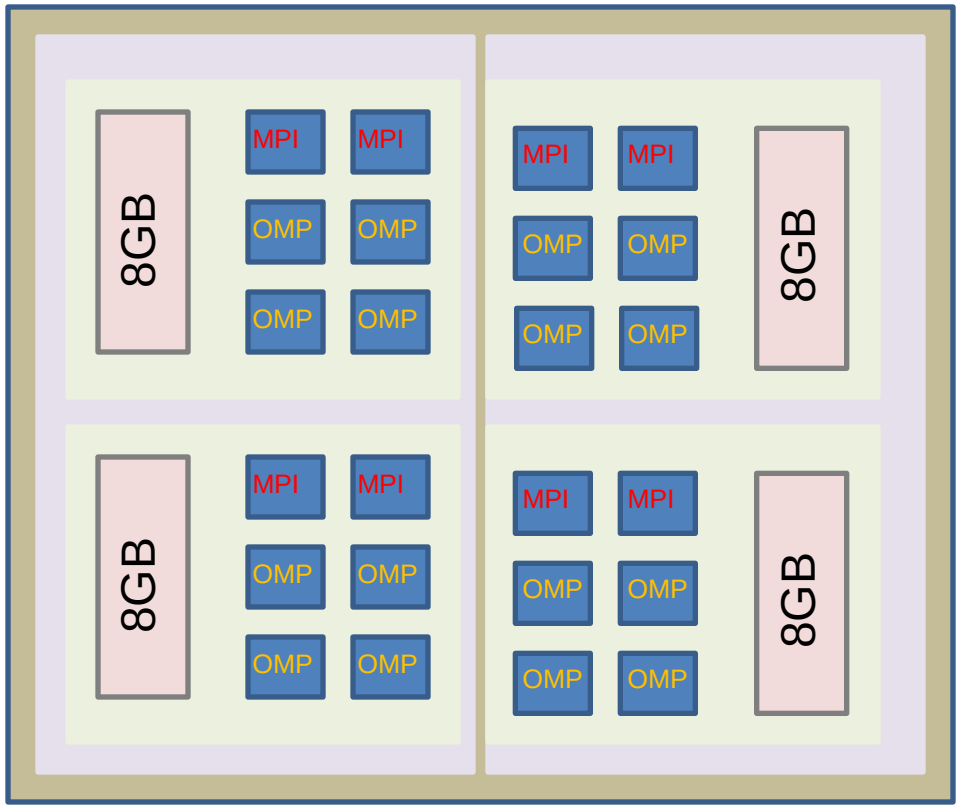


TABLE 1: Comparison of MPI and mixed-mode parallel on XT4h

MPI tasks	16	32	64	128
XT4h GMM N4	4.227	2.174	1.426	1.085
XT4h GMH N1t1	2.696	1.498	0.826	0.665
XT4h GMH N1t2	1.692	0.979	0.58	0.574
XT4h GMH N1t4	1.337	0.735	0.489	0.473

CHART 1: Effect of Open MP on time per step on XT4h

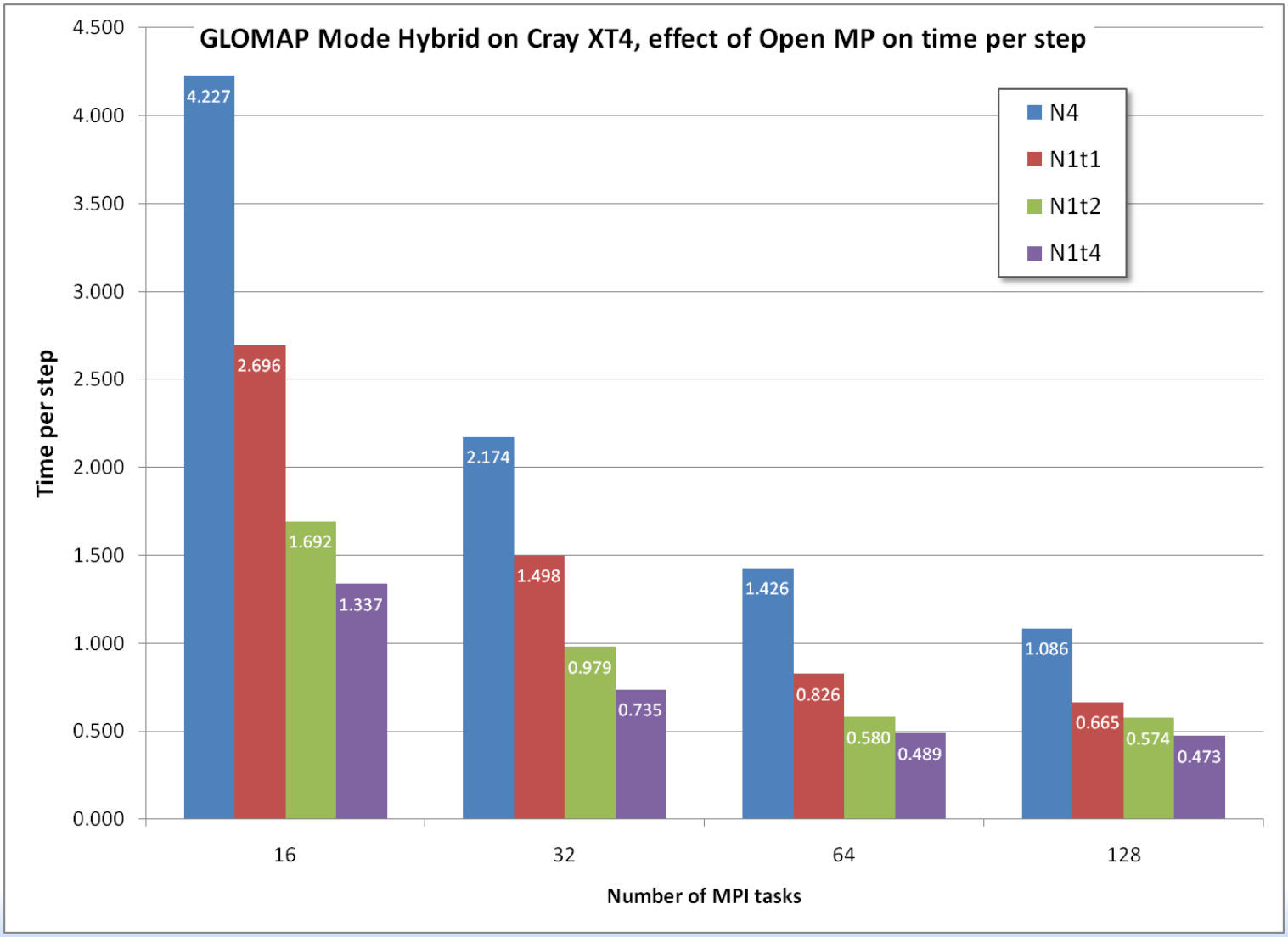


Chart 2 : Effect of Open MP on cost per step on XT4h

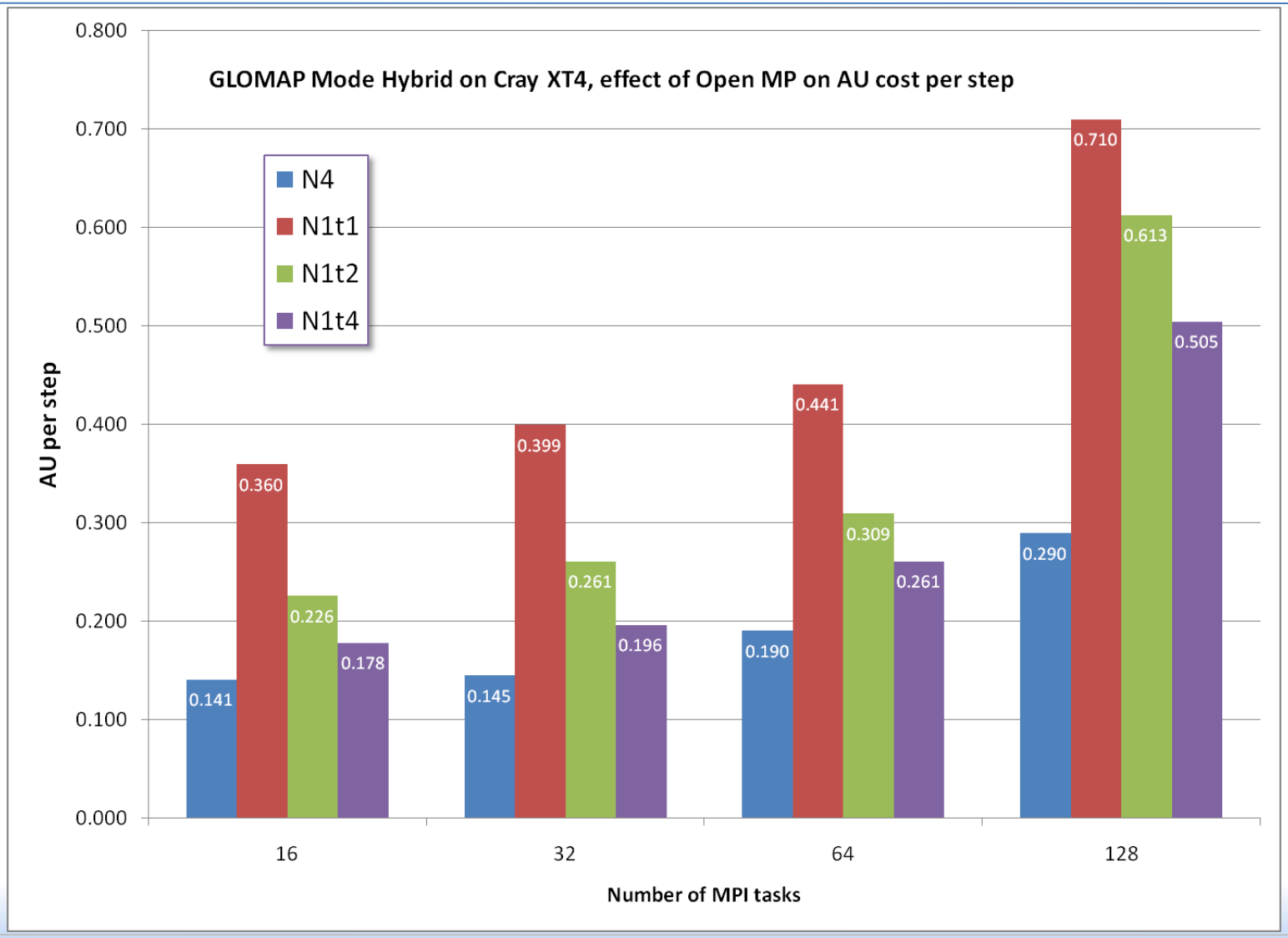


CHART 3 : XT4h : time per step with fully populated nodes

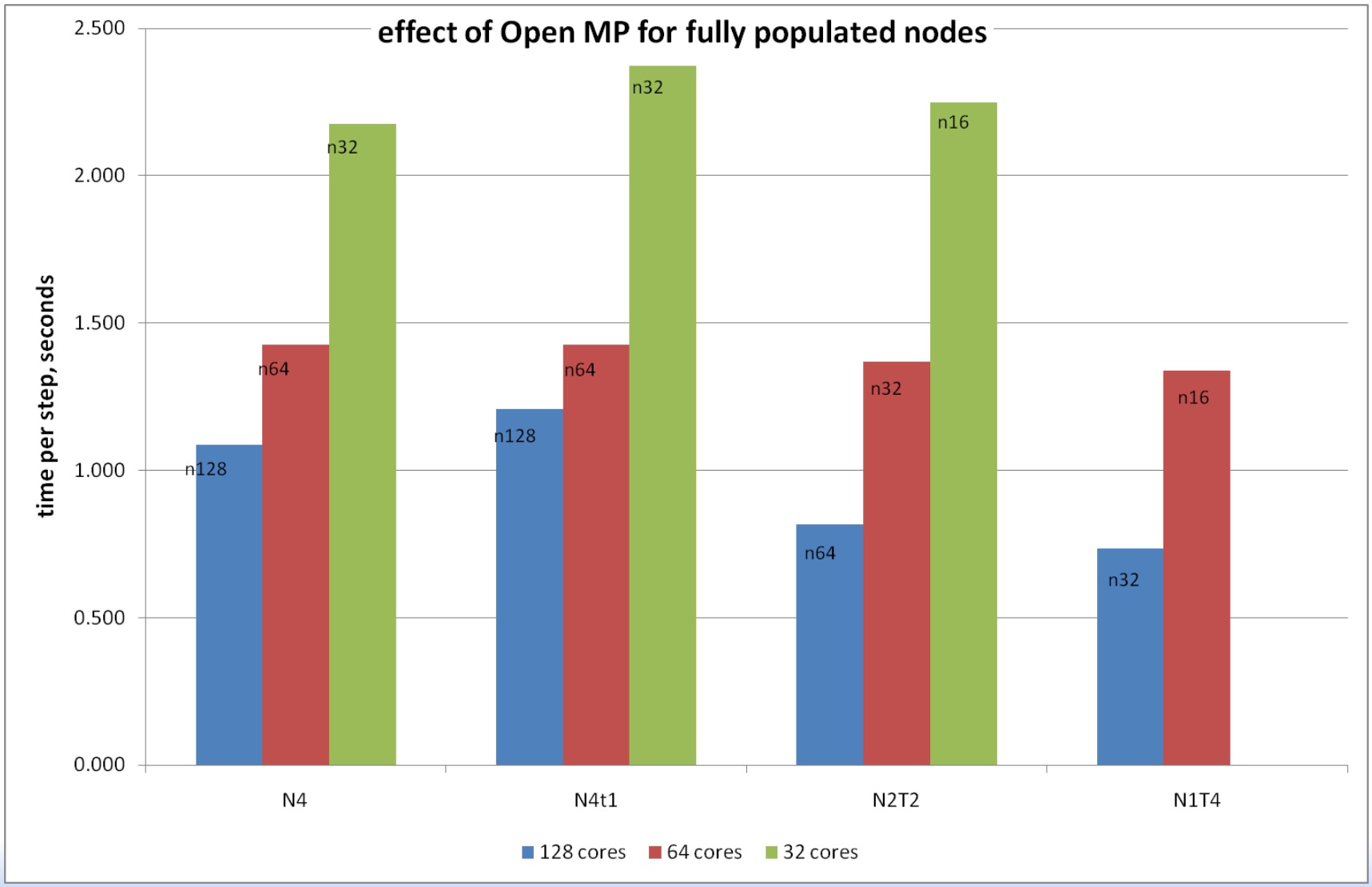
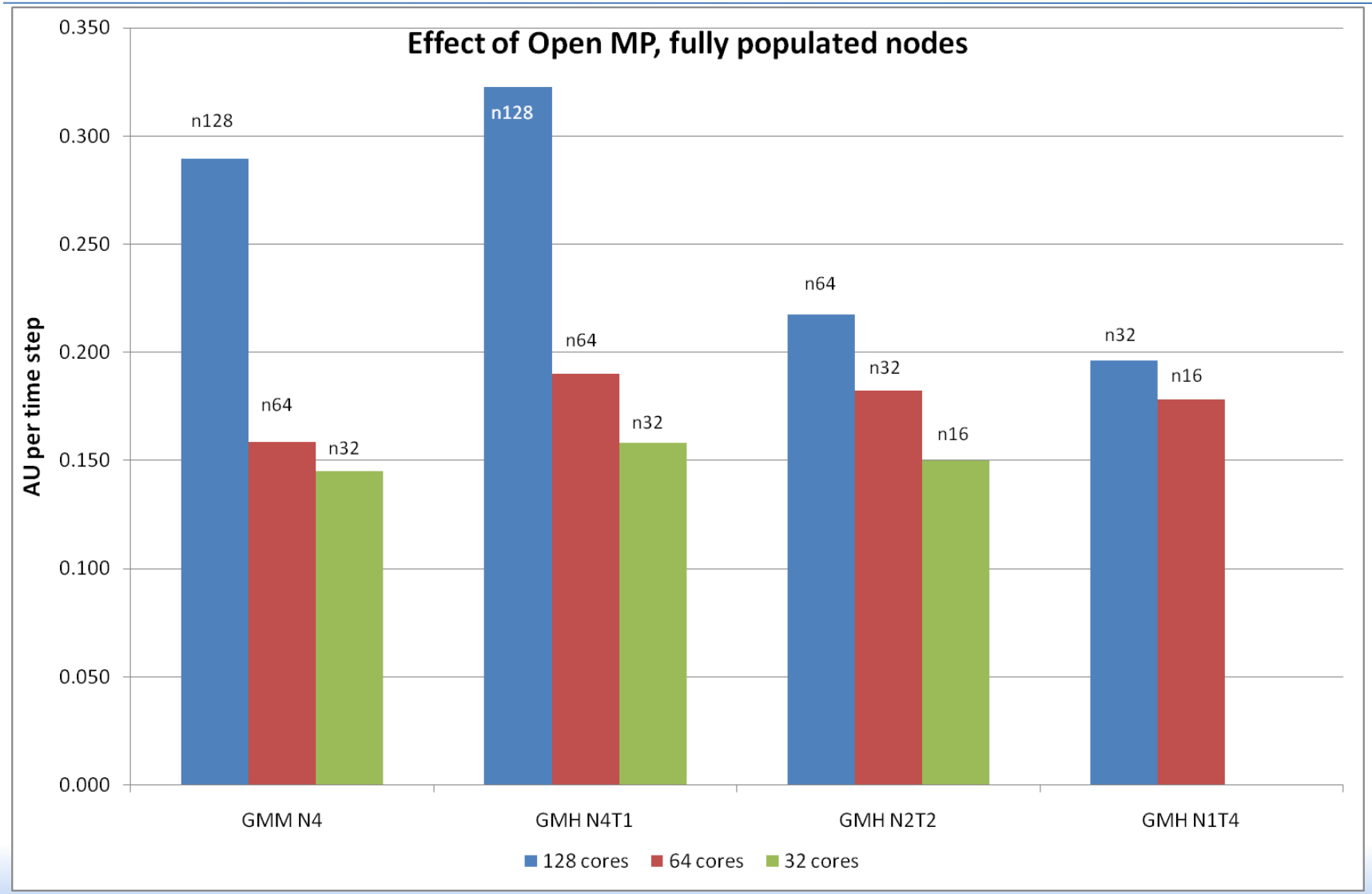


CHART 4 : XT4h : Cost as AU per step



Results XT6

- Two MPI configurations M32 and M64
- Various node use configurations
  - Pure MPI to see effect of the “powers of 2” and idle cores
  - Mixed-mode parallel operation to utilise the idle cores
- Notation used to distinguish configurations
  - *n*=MPI tasks, N=number of tasks per node, t=number of threads per MPI task
  - XT6 work S=MPI tasks per hex-core die and D the number of threads per MPI tasks
- Time per step and AU often presented on same axis
  - Useful to see together and are same order of magnitude

CHART 5 : MPI-only XT6, 64 MPI tasks, speed improvement from placement

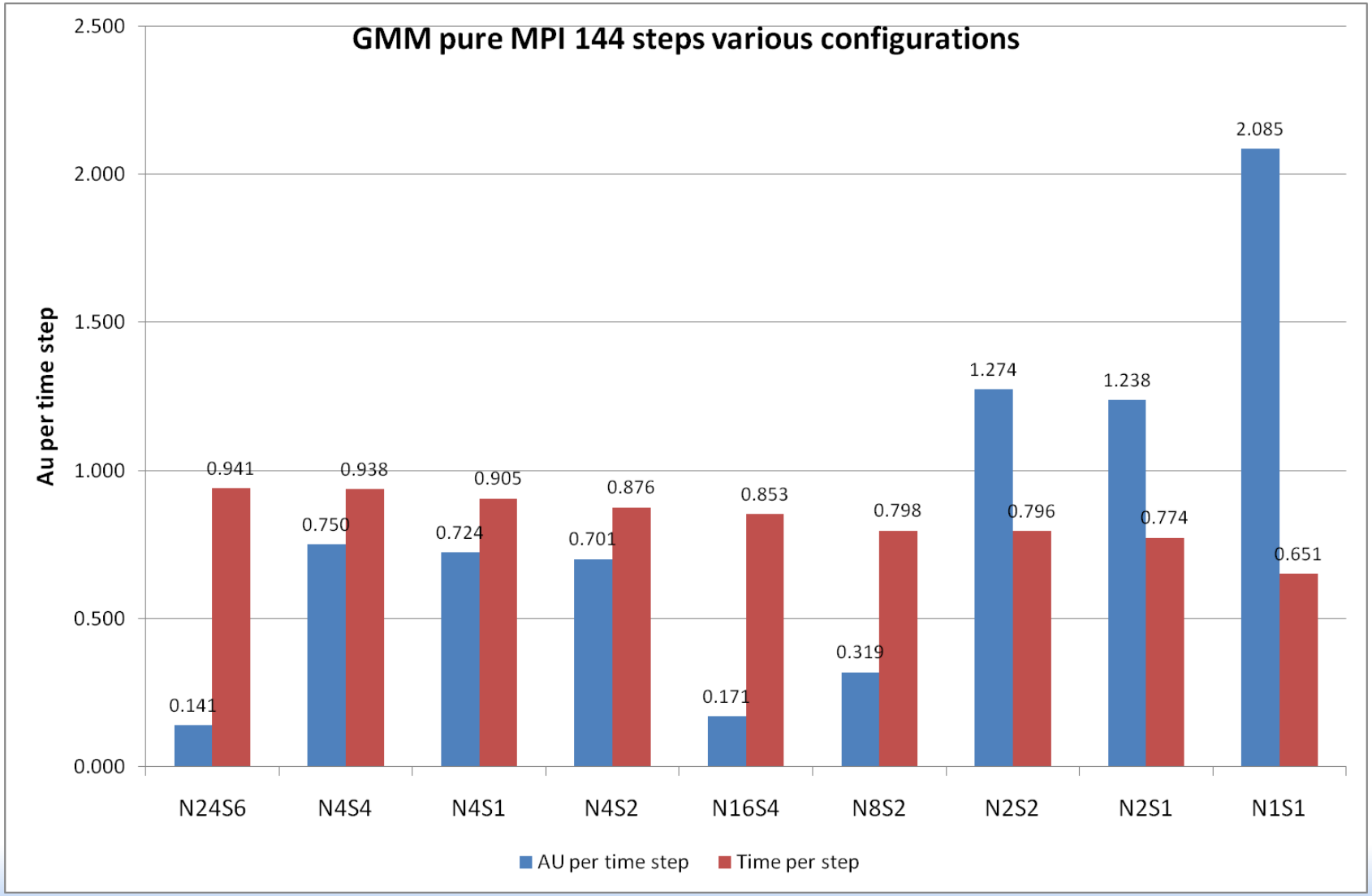


CHART 6 : Speed-up due to Open MP 64 MPI tasks

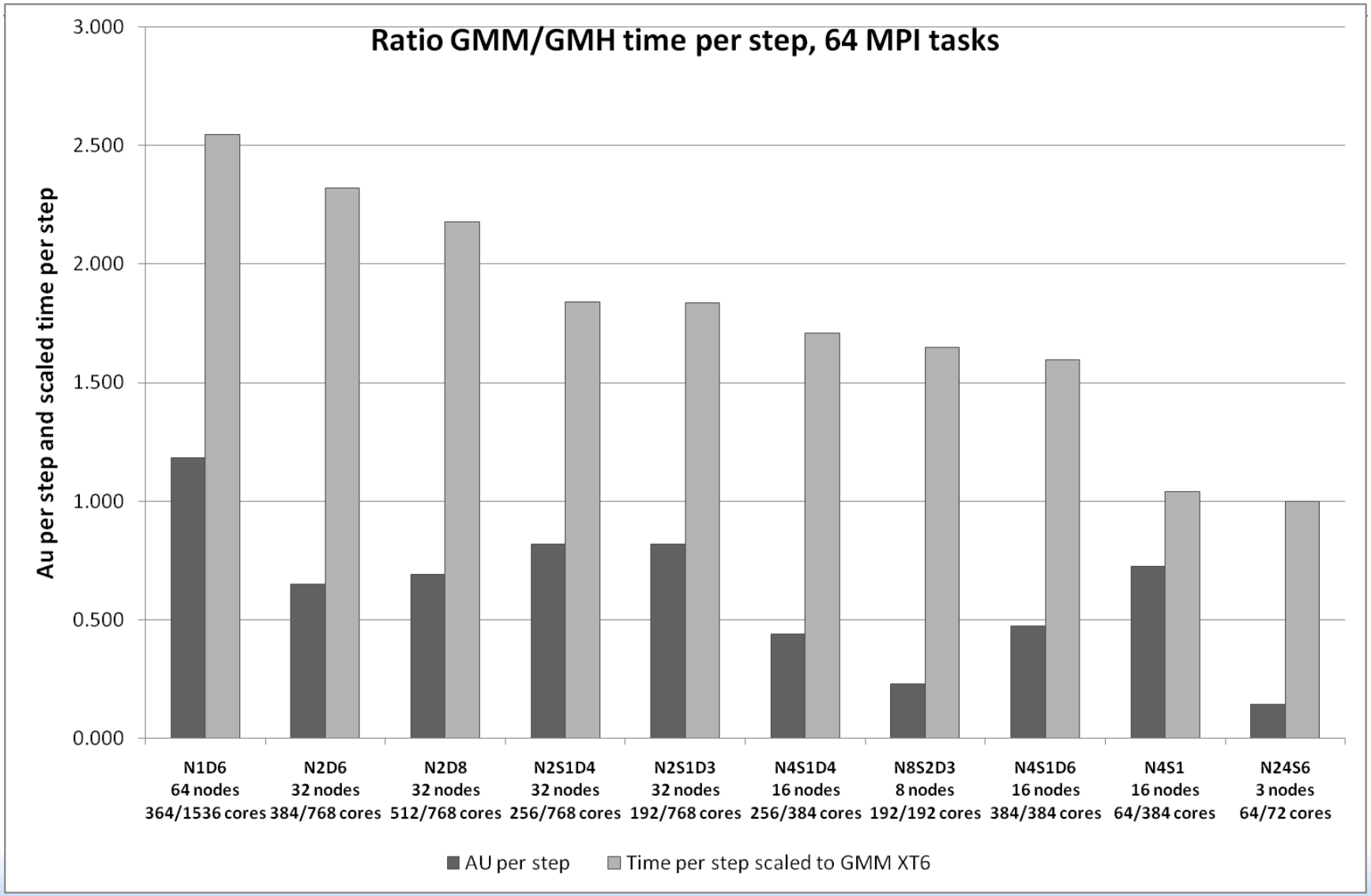


CHART 7 : mixed mode XT6, 64 MPI tasks

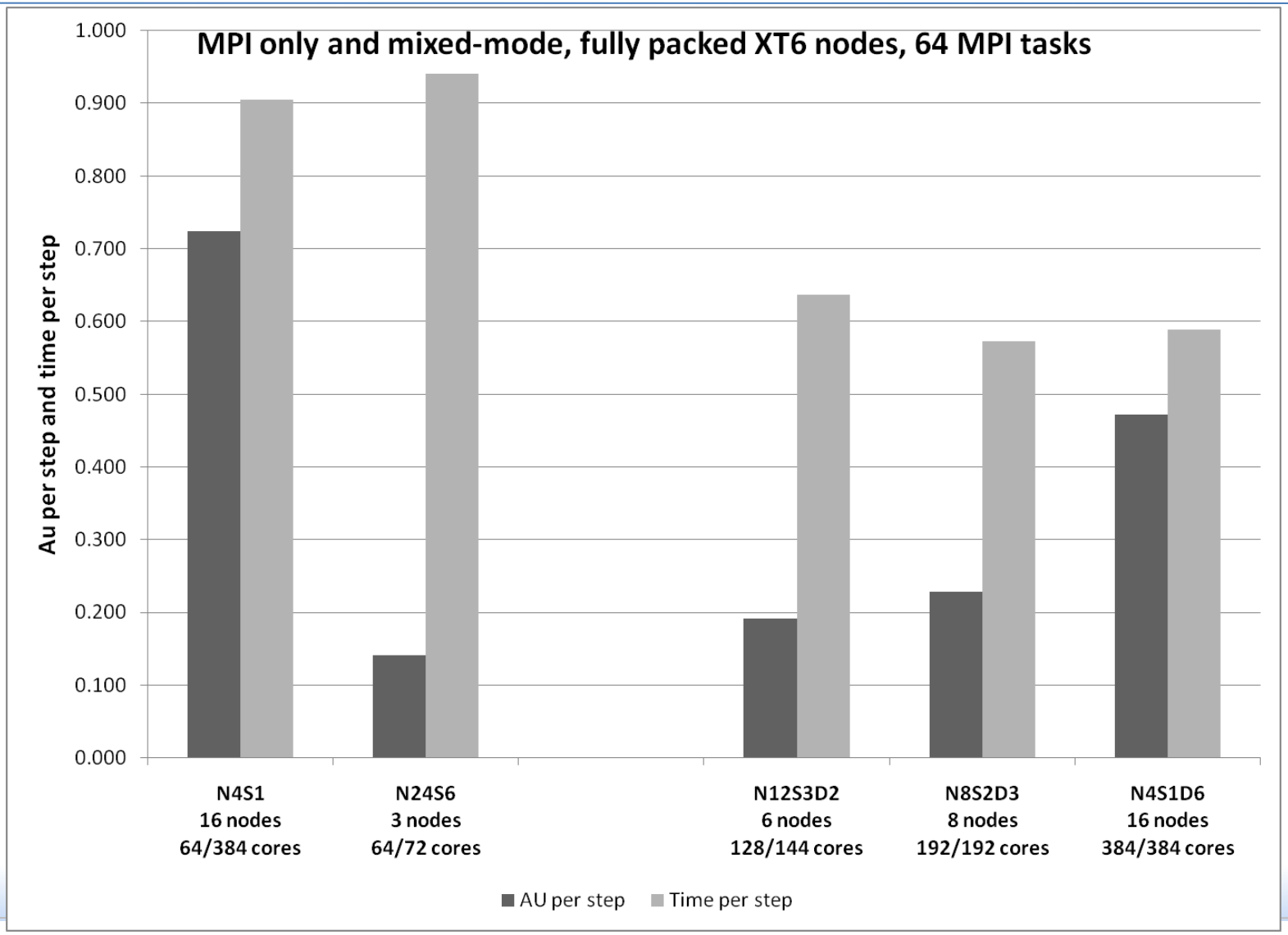
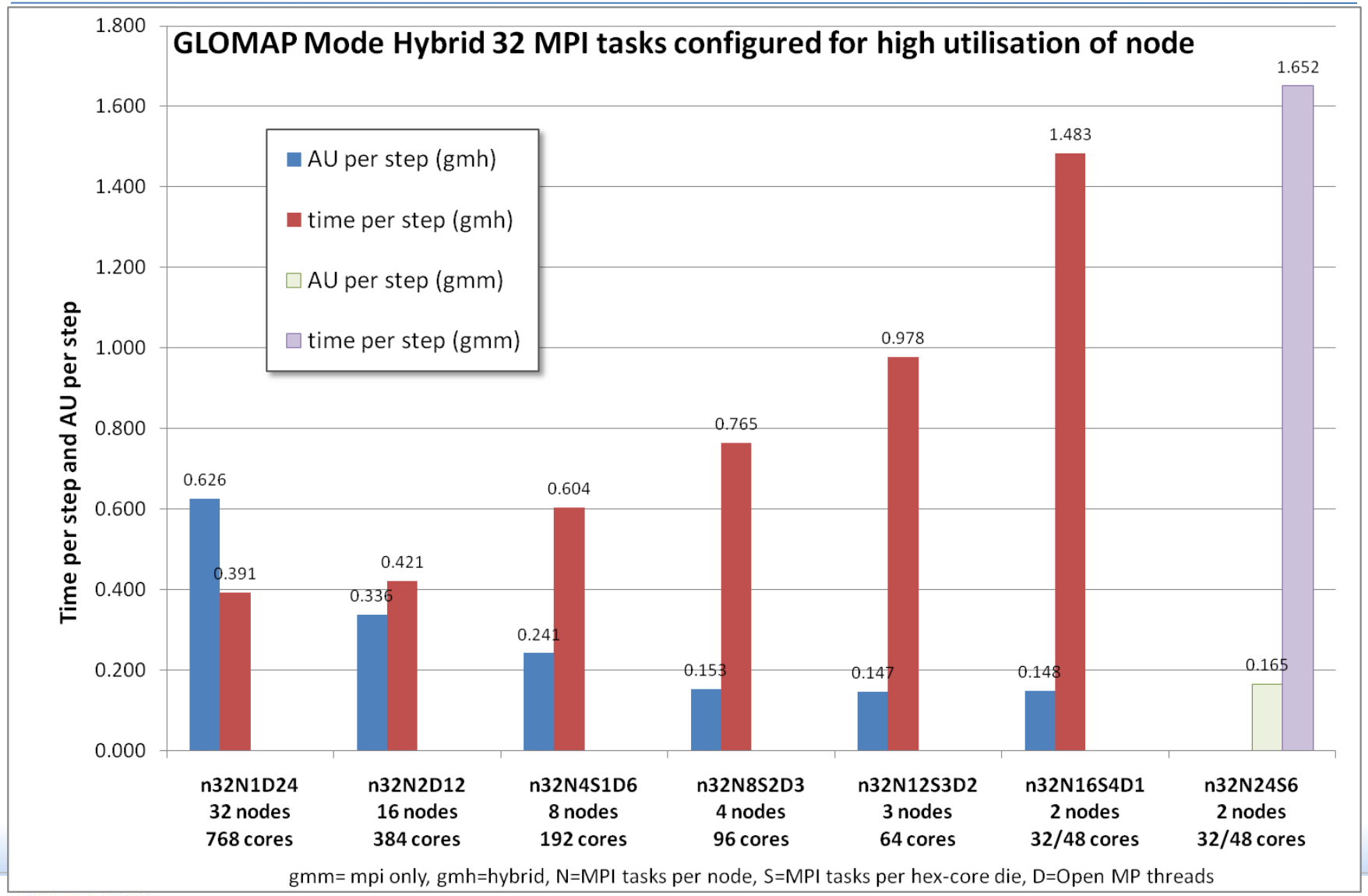


CHART 8 : Mixed-mode XT6, 32 MPI tasks



Open MP acceleration

- This is reasonable on the XT4h
- This is non-ideal on XT6
  - Should not expect to scale directly with number of threads
  - Loop limits are not well balanced for more than 8/4/2 threads
    - Depending on MPI decomposition
- There is X% of code where Open MP implemented
  - Only applied to the significant workload loops
  - Much larger effort would be required to fix every subroutine (330)

Summary

- GLOMAP Mode MPI has been analysed with a view to
- Open MP has been added to an MPI code
- The test case has been exercised on the Xt4 and XT6
- The results have been presented here and I hope you will agree that the following conclusions are valid…

Conclusions

- The cost of using more nodes is recovered almost fully
  - through the increased speed of each MPI task
  - by multi-threaded acceleration
  - by reducing the number of intra-node MPI tasks
- Placement of the tasks on a “many-core” system is critical to performance
- The shorter run-times with mixed-mode will allow more research to be done
- Fewer MPI tasks per node will allow higher resolution simulations
- GLOMAP Mode MPI will be ready for better use of XT6

Additional Observations

- Should look at extending places where OMP directives are used
- Debugging is a night mare
  - Need the developer on hand
- Examine the option “auto parallel” and how it interacts with Open MP